Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Jan 29th. 2019

Please note that assignments in this class may be submitted to

www.turnitin.com, a web-based anti-plagiarism system, for an evaluation of their originality.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Reading review**

**Scaling Big Data Mining Infrastructure:**

**The Twitter Experience**

This paper talks about Twitter’s infrastructure evolution and the capabilities development for data mining on ‘big data’.

So how Twitter develop like this? There are 3 major factors:

1. Now, we live in a world of data explosion.
2. The techniques of data analysis are increased, from simple to complex.
3. As open-source software, Twitter plays an important role in the ecosystem.

The big data mining cycle part told us as a data scientist, it is necessary to know that is available data and how they are organized before analyzing the data. And this part includes 4 methods:

1. Service architectures and logging
2. Exploratory data analysis
3. Data mining
4. Production and other considerations

Why should we consider about this?

The consideration of an academic researcher is only a small part of the complete data mining cycle. However, we should know that big data mining infrastructure have to consider all activities in the lifecycle, just use data mining algorithms are insufficient.

The schemas and beyond part introduce the importance to use right methods as a log.

And in the last part, we learned 3 main components of data mining solutions:

1. The raw data
2. The feature representation extracted from the data
3. The model or algorithm used to solve the problem.